2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) |978-1-6654-0392-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/TRUSTCOM50675.2020.00246

Trust and Believe – Should We? Evaluating the Trustworthiness of Twitter Users

1st Tanveer Khan
Network and Information Security Group
Tampere University
Tampere, Finland
tanveer.khan@tuni.fi

2nd Antonis Michalas
Network and Information Security Group
Tampere University
Tampere, Finland
antonios.michalas@tuni.fi

***Abstract*—Social networking and micro-blogging services, such as Twitter, play an important role in sharing digital information. Despite the popularity and usefulness of social media, they are regularly abused by corrupt users. One of these nefarious activities is so-called fake news – a virus that has been spreading rapidly thanks to the hospitable environment provided by social media platforms. The extensive spread of fake news is now becoming a major problem with far-reaching negative repercussions on both individuals and society. Hence, the identification of fake news on social media is a problem of utmost importance that has attracted the interest not only of the research community but most of the big players on both sides - such as Facebook, on the industry side, and political parties on the societal one. In this work, we create a model through which we hope to be able to offer a solution that will instill trust in social network communities. Our model analyses the behaviour of 50,000 politicians on Twitter and assigns an influence score for each evaluated user based on several collected and analysed features and attributes. Next, we classify political Twitter users as either trustworthy or untrustworthy using random forest and support vector machine classifiers. An active learning model has been used to classify any unlabeled ambiguous records from our dataset. Finally, to measure the performance of the proposed model, we used accuracy as the main evaluation metric.**

***Index Terms*—Credibility, Fake News, Influence Score, Sentiment Analysis, Trust, Twitter, Active Learning**

I. INTRODUCTION

With one-third of the world's population using some form of social media [61], it is evident that the popularity of social networking sites has rapidly increased in recent years. This has significantly changed the dynamics of communication across all age groups; the way we work, the way we live, the way we interact with other people and the way we share information have already changed drastically. Furthermore, social media enables sharing of important information with many people simultaneously, allowing users to reach a bigger audience.

While social media has its positive sides, it is also important to consider the flip side and properly evaluate its negative impacts. One of the latest negative effects of social media is the so-called fake news phenomenon. It has been proven that the massive distribution of fake news plays an important role in the success or failure of important events and causes [10], [11]. Apart from the dissemination and circulation of false information, social networks provide the ideal toolkit for corrupt users to perform a wide range of illegitimate actions such as spamming and political Astroturfing [7], [9].

Twitter, with around half a billion users, is one of the three most popular social media platforms. It generates on average 10,000 tweets per second (approximately 500 million tweets per day[1]) [47]. It is considered a valuable resource for government agencies, businesses, political parties, financial institutions, fundraising, and many other actors as it enables uncomplicated extraction and dissemination of important information.

A recent study [1] examined 10 million tweets generated by 700,000 different Twitter accounts and linked to 600 fake and conspiracy news sites. It identified clusters of Twitter accounts that linked back to these sites repeatedly, often in ways that seemed coordinated or even automated. In another study, it was found that 6.6 million tweets with fake news were distributed before the 2016 US elections. Different social and political events such as the 2016 US presidential election [15] were tainted by a growing number of fake news.

Global concern about the impact of fake news on our societies is on the rise. Hence, there is an immediate need for the design, implementation, and adoption of new systems and algorithms that are able to *identify* and *differentiate* between fake and real news. However, with the increase in the number of social media users[2], the quantity of generated content is increasing rapidly, which hinders the identification of fabricated stories [16] and prevents the identification of a significant amount of information that can potentially give rise to false rumours. Therefore, verifying the credibility of a tweet or assigning a score to users based on the information they have been sharing is a problem that has caught the interest of many academic and industrial researchers [17], [18], [20]–[25].

A. Our Contribution

In this work, we present a model for analysing Twitter users that assigns a score calculated based on their social profiles,

This research has received funding from the EU research projects ASCLEPIOS (No. 826093) and CYBELE (No 825355).

[1]https://www.omnicoreagency.com/twitter-statistics/

[2]In 2018, an estimated 2.65 billion people were using social media worldwide, a number projected to increase to almost 3.1 billion in 2021 [61].

978-1-6654-0392-4/20/$31.00 ©2020 IEEE
DOI 10.1109/TrustCom50675.2020.00246

tweet credibility and h-index score (i.e. retweets and likes). Users with a higher score are not only considered to be more influential but their tweets are also given greater credibility. Our main contribution can be summarised as follows:

- First, we generated a dataset of 50,000 Twitter users. For each user, we created a unique profile containing 19 features (discussed in Section III). Our dataset contained only users whose tweets are public and who have friends and followers.
- For each of the analysed users, we calculated their Social Reputation score (Section III-B), an h-Index Score (Section III-B), a Sentiment Score (Section III-B), Tweet Credibility (Section III-B) and an Influence Score III-C.
- Furthermore, we classified each Twitter user account as either trustworthy or untrustworthy. A trustworthy or untrustworthy flag was assigned to each user based on their social reputation, tweet credibility, the sentiment score of a tweet and H-index score of re-tweets and likes, as well as an influence score.
- To classify a large pool of unlabeled data, we used an active learning model (a semi-supervised learning algorithm) – a technique ideal for a situation in which unlabeled data is abundant but manual labeling is expensive [63], [67].
- We measured the performance of our model by using the accuracy metric. This metric measures the percentage of correctly predicted Twitter users (trustworthy and untrustworthy).

We hope that this work will inspire others to perform further research on this emerging problem while at the same time kick-starting a period of greater trust on social media through sustained collaboration between humans and machines.

B. Organisation

The rest of this paper is organised as follows: In Section II related work is discussed followed by Section III in which we discuss in detail our proposed approach. The active learning approach and types of classifiers used are discussed in Section IV. Section V features the experimental results and model evaluation and presents the data collection and experimental results of our model. Finally, in Section VI, we conclude the paper.

II. RELATED WORK

Twitter is considered one of the top Online Social Networks (OSNs) that provide a fertile environment for a variety of research purposes. Compared to other popular OSNs, Twitter gains significantly more attention in the research community due to its open policy on data sharing and distinctive features [4]. In 2011, the network had about 175 million unique accounts [27], a figure that has grown to an estimated 1.3 billion[3], making it one of the most popular social media platforms.

Even though openness and vulnerability are two separate issues, there have been many cases where malicious users have taken advantage of Twitter's openness and managed to exploit the service in several ways (e.g. political Astroturfing, spammers sending unsolicited messages, posting malicious links, etc.).

Despite the important negative impact that the distribution of fake news has on our society, only a handful of techniques for identifying fake news on social media have been proposed [4], [7], [9], [30], [31]. One of the most popular and promising ideas is to evaluate Twitter users and assign them a credit/reputation score.

Authors in [7] elaborated on the idea that posting duplicate tweets should affect the reputation score of a user since this is a behaviour that legitimate users typically do not engage in. Therefore, posting the same tweet several times would have a negative effect on the user's overall credit score. The authors calculated the edit distance to detect duplication between two tweets posted from the same account. Furthermore, the staggering quantities of exchanged messages and information on Twitter have been exploited by users to hijack trending topics [8]. This is a technique used to send unsolicited messages to legitimate users. Additionally, there are Twitter accounts whose only purpose is to artificially boost the popularity of a hashtag with the main aim of increasing its popularity and ultimately making the underlying topic a trend. One BBC report mentioned that £150 was paid on Twitter users to increase the popularity of a hashtag and make it a trend[4].

To tackle these problems, researchers have used different ways to assess the trustworthiness of tweets and assign an overall rank to users [31]. Castillo *et al.* [35] measured the credibility of tweets (news topics) based on Twitter features. More precisely, an automated classification technique to detect news from conversational topics was used. Alex Hai Wang [7] used followers and friends parameters to calculate the reputation score, which further aided user classification (i.e. to detect spammers). Additionally, Saito and Masuda [60] considered these metrics while assigning a rank to Twitter users. In [36], the authors analysed tweets relevant to the Mumbai attacks[5]. Their analysis showed that most information providers were unknown while the reputation of the others (based on number of followers) was very low. In another study [37] that looked at the same event, an information retrieval technique and machine learning algorithm found that only 17% of the tweets related to the underlying attacks were credible.

Gilani *et al.* [43] found that compared to normal users, bots and fake accounts use a large number of external links in their tweets. Hence, analysing other Twitter features such as URLs is of paramount importance for correctly evaluating the overall credibility of a user. While Twitter has built tools to filter out such URLs, there are several masking techniques that can effectively bypass Twitter's safeguards.

[3]https://www.brandwatch.com/blog/twitter-stats-and-statistics/

[4]https://www.bbc.com/news/blogs-trending-43218939

[5]https://www.theguardian.com/world/blog/2011/jul/13/mumbai-blasts

In this work, we evaluate the trustworthiness and credibility of users [5], [6] by analysing a wide range of features (see Table I). Compared to other works in the area, our model sifts through a plethora of factors that represent signs of possible malicious behaviours and makes honest, fair and precise judgments about the credibility of users with the main aim of engendering community trust.

III. Methodology

In this section, we discuss our models and main algorithms for calculating the influence score of a user. Our first goal is to help users identify possible information about a political Twitter user by taking into consideration the influence score that results from a proper run of our algorithms. Secondly, political Twitter users are classified either as trustworthy or untrustworthy users based on social reputation, tweet credibility, sentiment score, h-index score and influence score. All those accounts with abusive and harassing tweets, a low social reputation, h-index and influence score are grouped into untrustworthy users while those who are more reputable among users with a high h-index score and more credible tweets as well as high influence score are grouped into trustworthy users. In the rest of section, we will talk about how to calculate the influence score of a Twitter user. The influence score is calculated by considering both the context and content (tweets) of Twitter accounts. In evaluating a user we take into consideration only the Twitter features that can be extracted using Twitter API. Then, we use the outcome of that evaluation and derive more features that help us to provide a more well-rounded and fair evaluation (Section III-B). Figure 1 illustrates the main factors we used to calculate user influence scores.

A. Twitter Feature Extraction

We now describe in detail all the basic features extracted from Twitter and their importance in the process of assigning a score to each user.

The key step in assigning a score to Twitter users is to extract the features linked to their accounts. The features can be specific to user accounts such as the number of followers and friends or it can be specific to a user's tweet such as the number of likes, retweets, URLs, etc. In our model, we used these as well as additional features on both a user and content level.

The features used to assign an influence score as well as the relevant notation used throughout the paper is given in table I.

Following or Friending: Following or friending are user-level features indicating that a Twitter user has subscribed to the updates of another user (i.e. following another user) [26]. Following users who are not part of one's interpersonal network results in a large amount of novel information. One of the important indicators for calculating the influence score for Twitter users is the $followers/following$ ratio. The $follower/following$ ratio compares the number of u_i's subscribers to the number of users u_i is following. Users are more interested in updates if the $follower/following$ ratio is high [27]. In our model, we use $friends$ as one of the indicators when assigning a social reputation score to a user.

TABLE I: Selected Attributes for Calculating the Influence Score

Notation	Description
$SN(u_i)$:	User screen name
$Id(u_i)$:	User's unique ID
$R(u_i)$:	User influence score
$N_T(u_i)$:	Total number of tweets
$N_{+ve}(u_i)$:	Number of neutral tweets
$P_{+ve}(u_i)$:	Number of positive tweets
$N_{-ve}(u_i)$:	Number of negative tweets
$S(u_i)$:	User status
$L(u_i)$:	User list count
$R_{hindex}(u_i)$:	User retweet h-index
$L_{hindex}(u_i)$:	User like h-index
$C_s(u_i)$:	User sentiment score
$Twt_{cr}(u_i)$:	User tweet credibility
$R_s(u_i)$:	Social reputation score of the user
$F_{fol}(u_i)$:	Number of user followers
$F(u_i)$:	Number of user friends
$M(u_i)$:	Number of user mentions
$U_R(u_i)$:	User tweets containing $URLs$
$R_R(u_i)$:	User retweet ratio i
$L_R(u_i)$:	User liked ratio
$O_R(u_i)$:	Original content ratio of the user
$H(u_i)$:	User hashtag ratio
$U(u_i)$:	User URL ratio
$M_R(u_i)$:	User mention ratio
R_n:	Number of retweets
L_n:	Number of likes
I_t:	Index number of tweets

Number of Followers: Number of followers is another user-level feature that shows the number of people interested in the tweets of a specific user u_i. As discussed in [57], number of followers is one of the most important parameters for measuring user influence. The more followers a Twitter user has, the more influential [58] a user is. Preussler *et al.* [59] correlate the number of followers with the reputation of the Twitter user. According to their study, as the number of followers increases, the importance/credibility of the underlying user also increases. Based on these studies, we also considered the number of followers as an important parameter and used it as an input to calculate the social reputation of the user.

Number of Retweets: A tweet is considered important when it receives many positive reactions from other accounts. The reactions may be in the form of likes or retweets. Retweets act as a form of endorsement, allowing Twitter users to

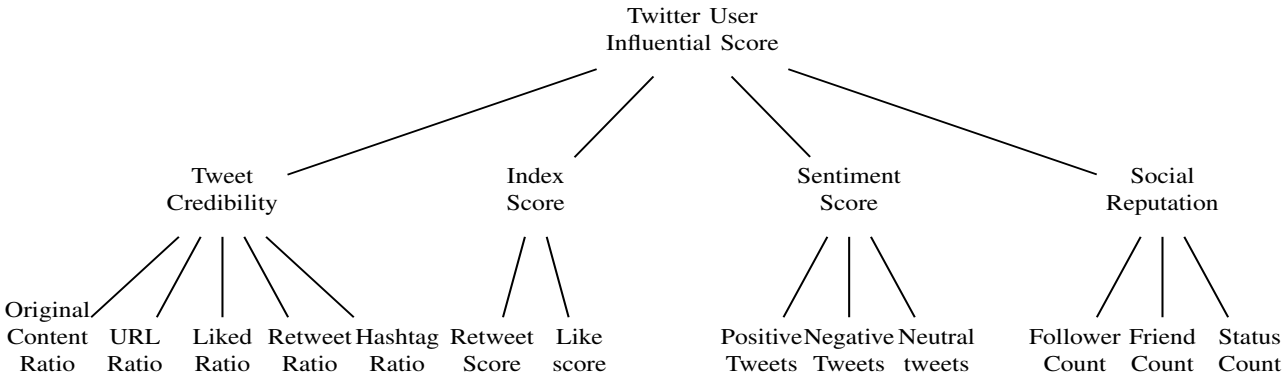


Fig. 1: Twitter User Influence Score Calculation

forward the content generated by other users, thus raising the content's visibility. Hence it is a way of promoting a topic and is associated with the reputation of the user [28]. Since retweeting is linked with popular topics and directly affects the reputation of a user, it is a key parameter for identifying possible fake account holders. As described in [43], bots or fake accounts depend more on retweeting existing content rather than posting new tweets. In our model, we consider the number of retweets as one of the main parameters for assigning an influence score to user accounts. To do so, we calculate the number of times a tweet is retweeted. Additionally, we calculate the total number of tweets for each user. The total number of tweets of user u_i is denoted by $N_T(u_i)$. Finally, we calculate the retweet ratio (using Twitter grader) for each tweet by considering a tweet that is retweeted divided by the total number of tweets, given in equation (1).

$$R_R(u_i) = \frac{Retweets}{N_T(u_i)} \tag{1}$$

Likes: The number of likes is a reasonable proxy for evaluating the quality of a tweet. The authors in [41] showed that humans receive more likes per tweet when compared with bots. In [42], the authors used likes as one of the metrics for classifying Twitter accounts as human or automated. As mentioned in [47], if a specific tweet receives a large number of likes it can be safely concluded that other users are interested in the tweets of the underlying user. Based on this observation, we calculate the liked ratio by using the number of likes for each tweet and dividing it by the total number of tweets for the corresponding user as shown in equation (2).

$$l(u_i) = \frac{Liked\ tweets}{N_T(u_i)} \tag{2}$$

URLs: A URL is a content level feature that some users include in their tweets [48]. Since tweets are limited to a maximum of 280 characters, it is common that users cannot add important information to their tweets. To overcome this issue, tweets are commonly populated with URLs pointing to a resource where more information can be found. In our model, we consider the URL as an independent variable for engagement measurements [49]. We count the tweets that include a URL and calculate the URL ratio by considering the total number of tweets containing URLs over the total number of tweets as given in equation (3).

$$U(u_i) = \frac{Tweets\ with\ URLs}{N_T(u_i)} \tag{3}$$

Listed Count: In Twitter, a user has the option to form several groups by creating lists of different users[6] (e.g. competitors, followers, colleagues, etc.). Twitter lists are mostly used to keep track of the most influential people[7]. The simplest way to measure the influence of a Twitter user is by looking at the number of lists that the user is included on. Being a member of a large number of lists is an indicator that this user is considered important by others. Based on this assumption, in our model, we also consider the number of lists that each user belongs to.

Status Counts: Compared to the other popular OSNs, Twitter is considered to be a service that is *less* social[8]. This is mainly due to the large number of inactive users or users who show low motivation in engaging in an online discussion. Twitter announced a new feature "Status availability", that verifies the status of a user[9]. To this end, during the calculation of user influence scores, we also consider how active they are by measuring how often a user performs a new activity[10].

Mention by Others: A mention within a tweet contains another person's username anywhere in the body of the tweet. Due to the fact that mentions indicate the inclusion of a user in conversations, tracking Twitter mentions is one of the most effective ways to measure the presence of a user in the network. The retweet and mention ratio is calculated by Isabel and Christian [27] to judge the reaction from other users to a user tweet. In addition, these two parameters are also used in

[6]https://help.twitter.com/en/using-twitter/twitter-lists
[7]https://www.postplanner.com/how-to-use-twitter-lists-to-always-be-engaging/
[8]https://econsultancy.com/twitter-isn-t-very-social-study/
[9]https://www.pocket-lint.com/apps/news/twitter/
146714-this-is-what-twitter-s-new-online-indicators-and-status-updates-look-like
[10]https://sysomos.com/inside-twitter/most-active-twitter-user-data/

Twitter Grader (an online tool) to assign a score to a Twitter user. In our model, we use the retweet and mention ratio along with other indicators to check how influential a Twitter user is. The mention ratio can be calculated using equation (4).

$$M_R(u_i) = \frac{Tweets\ with\ mentions}{N_T(u_i)} \tag{4}$$

Original Content Ratio: It has been observed that most Twitter users retweet posts by others [27] instead of posting original tweets. As a result, Twitter has become a pool of constantly updating information streams. For users with high influence in the network, the best strategy is to use the $30/30/30$ rule: 30% retweets, 30% original content, and 30% engagement. With this in mind, in our model we are looking for original user tweets and add them to their influence score. We calculate the original content ratio by extracting retweet posts by others from the total tweets of users as given in equation (5).

$$O_R(u_i) = \frac{N_T(u_i) - Retweeted\ posts}{N_T(u_i)} \tag{5}$$

B. Derived Features for Twitter Users

In this section, we elaborate on the extraction of extra features after the consideration and evaluation of basic ones. Additionally, we discuss the sentiment analysis technique used to analyse user tweets.

By using the basic features described earlier, we calculated the following parameters for each user:

- Social reputation of the user;
- H-index score based on likes and retweets;
- Sentiment score;
- Tweet credibility;
- Influence score.

Social Reputation of the User: The main factor for calculating the social reputation $R_s(u_i)$ of a user u_i is the number of users that are interested in u_i's updates. Hence, u_i's social reputation is based on number of followers, friends and statuses [7], [27].

$$R_s(u_i) = \log((1 + F_{fol}(u_i)) \cdot (1 + F_{fol}(u_i))) + \log(1 + S(u_i)) - \log((1 + F(u_i)) \tag{6}$$

In equation (6), $R_s(u_i)$ is directly proportional to $F_{fol}(u_i)$ and $S(u_i)$. Based on several studies [7], [27], [47], the $R_s(u_i)$ of a user is more dependent on $F_{fol}(u_i)$ and that is the reason we give importance to $F_{fol}(u_i)$ in comparison to $S(u_i)$ and $F(u_i)$. If a user (u_i) has a large number of followers then it is evident that the user is more reputed in his network. In addition, if a u_i is more active in updating his/her status then there are more chances that the tweets from u_i receive more likes and get retweeted. As $F_{fol}(u_i)$ and $S(u_i)$ increase, u_i's social reputation $R_s(u_i)$ also increases and vice versa. Furthermore, if a user has less $F_{fol}(u_i)$ in comparison to $F(u_i)$ then obviously the $R_s(u_i)$ of a user is small. As can be seen from equation (6), there is an inverse relation between $R_s(u_i)$ and $F(u_i)$.

H-Index Score: The h-index score is most commonly used to measure the productivity and impact of a scholar or scientist in the research community. It is based on the number of publications as well as the number of citations for each publication[11]. In our work, we use the h-index score for a more accurate calculation of user influence scores. The h-index score of a Twitter user is calculated considering the number of likes and retweets for each tweet. To find the h-index score[12], we sort the tweets based on the number of likes and retweets (in decreasing order).

Algorithm 1 describes the main steps for calculating the h-index score of a Twitter user based on the *number of retweets* and *likes*.

Algorithm 1 Calculating a $User_i$ h-index score for retweets and likes

```
procedure H-INDEX SCORE(hindex(u_i))
    Arrange R_n/L_n for each tweet of the user in decreasing order
    for I_t in list: do
        if R_n/L_n of a tweet < I_t then
            return I_t
        end if
    end for
    return number of tweets
end procedure
```

$R_{hindex}(u_i)$ and $L_{hindex}(u_i)$ are novel features used to measure the relative importance of a user on Twitter. A tweet that has been retweeted many times and liked by many users is considered to be attractive to readers [19], [47]. For this reason, we use $R_{hindex}(u_i)$ and $L_{hindex}(u_i)$ to measure the influence of a Twitter user. The higher the $R_{hindex}(u_i)$ and $L_{hindex}(u_i)$ score of a Twitter user, the higher is that user's influence.

Credibility of Twitter Users: Credibility refers to believability [35], which requires reasonable grounds for being believed. Twitter user credibility can be assessed by using the information available on the Twitter platform. In our research work, we use both the sentiment score and tweet credibility in identifying credible Twitter users.

Sentiment Score: It has been observed that OSNs are a breeding ground for the distribution of fake news where even individual Twitter posts can have a significant impact [45] that will affect the outcome of an event.

With this in mind, we used sentiment analysis and the TextBlob [44] library to analyse recent tweets with the main aim of identifying certain attitudes/patterns that can lead us to the identification of credible news. The sentiment analysis returned a score using polarity values ranging from 1 to -1 and helps in tweet classification. We classify the collected tweets as *(1)* Positive P_{+ve} *(2)* Neutral N_{+ve}, and *(3)* Negative N_{-ve}

[11]https://www.researchgate.net/post/How_to_calculate_h_index_for_an_author

[12]https://gallery.azure.ai/Notebook/Computing-Influence-Score-for-Twitter-Users-1

based on the number of positive, neutral and negative words in a tweet with $P_{+ve}(u_i)$ being the most credible tweets followed by the neutral tweets $N_{+ve}(u_i)$ and then the least credible tweets $N_{-ve}(u_i)$. According to Morozov *et al.* [46], the least credible tweets are associated with negative social events. They have more negative words and opinions, while credible tweets have more positive opinions and words.

After classification, based on previous tweets, we assign a sentiment score to each user (u_i) [47] using the following equation:

$$C_s(u_i) = \frac{\sum N_{+ve} + \sum P_{+ve}}{\sum N_{+ve} + \sum P_{+ve} + \sum N_{-ve}} \tag{7}$$

Tweet Credibility: Donovan [51] focused on finding the best indicators for credibility. According to these results, the best indicators for tweet credibility are URLs, mentions, retweets and length. Gupta *et al.* [37] ranked tweets based on tweet credibility. The parameters used as input for the ranking algorithm are total unique users, tweets, tweets with URLs, single tweets, retweets, trending topics, start and end date. Based on the existing literature, we compute the credibility of tweets by considering $R_R(u_i)$, $L_R(u_i)$, $H_R(u_i)$, $U(u_i)$ and $O_R(u_i)$:

Based on the above parameters, we measure tweet credibility by using (equation (8)).

$$Twt_{cr}(u_i) = \frac{((R_R(u_i)+L_R(u_i)+H(u_i)+U(u_i))}{4} \cdot O_R(u_i)) \tag{8}$$

For calculating the credibility of tweets, first we extract the $O_R(u_i)$ published by a Twitter user as a tweet. Then we consider the number of times this tweet is $R_R(u_i)$ and $L_R(u_i)$ by others. In addition, we also consider $H(u_i)$ and $U(u_i)$ as they are the functions that can be used for user engagement. Since these four parameters, $R_R(u_i)$, $L_R(u_i)$, $H(u_i)$ and $U(u_i)$, are linked with $O_R(u_i)$, we start by calculating the average of these four parameters and then multiply that result by $O_R(u_i)$. Based on these parameters, we calculate the credibility of tweets as given in equation (8).

C. Influence Score

The influence score of a Twitter user is calculated based on the evaluation of *both* content and context features. More precisely, we consider the following parameters described earlier: $R_s(u_i)$, $C_s(u_i)$, $Twt_{cr}(u_i)$ and $hindex(u_i)$. After calculating the values of all of these features, we use them as input for Algorithm 2 line 7. which calculates the influence score for the underlying Twitter user.

Equation Formulation: To find out how influential a Twitter user is, researchers have taken into consideration one, two or more of the following characteristics:

- Weight-age assigned to their tweets and impact [47];
- Credibility of the tweets [47], [51];
- Social reputation of the Twitter user [52];
- Level of activity, involvement in follow-up and discussions and the ability to propose new ideas [53].

An influential Twitter user must be highly active (e.g. able to start new discussions, have ideas that impact other users' behaviors, etc.). Additionally, the user's tweets must be credible, relevant and highly influential (i.e. liked and retweeted by a large number of other users). If the tweets of highly influential users are credible and the polarity of their tweet content is positive, they are highly acknowledged and recognised by the community. In short, for a Twitter user to be considered influential, we combine the efforts of [47], [51]–[53] and calculate the influence score through Algorithm 2 line 7.

Algorithm 2 shows the steps we follow to calculate the influence score of a user.

Algorithm 2 Calculating the User u_i Influence Score

1: **procedure** INFLUENCE SCORE($R_s(u_i)$)

2: For i^{th} user u_i

3: Calculate R_{hindex} and L_{hindex} of User u_i by using Algorithm 1

4: Calculate Sentiment Score of User u_i

$$C_s(u_i) = \frac{\sum N_{+ve}+\sum P_{+ve}}{\sum N_{+ve}+\sum P_{+ve}+\sum N_{-ve}}$$

5: Calculate Tweet Credibility of User u_i

$$Twt_{cr}(u_i) = [\frac{R_R(u_i)+L_R(u_i)+H_R(u_i)+U_R(u_i)}{4}] \cdot O_R(u_i))$$

6: Calculate Social Reputation of User u_i

$$R_s(u_i) = \log((1 + F_{fol}(u_i)) \cdot (1 + F_{fol}(u_i))) + \log(1 + S(u_i)) - \log((1 + F(u_i))$$

7: Compute Influence Score of User u_i

$$R(u_i) = \frac{C_s(u_i)+Twt_{cr}(u_i)+R_s(u_i)+R_{hindex}(u_i)+L_{hindex}(u_i)}{5}$$

8: **end procedure**

D. Parameter Selection and Comparison with Previous Models

The parameters used for calculating the influence score are based on an extensive study of the existing literature. The selected parameters are used for detection purposes [54]–[56], assigning a score [31] or for classification purposes [41]. We used all these parameters to assign an influence score to users. Table II provides an overview of comparisons between existing models based on feature selection.

IV. Active learning and ML Models

In the existing literature, the classification of Twitter users is primarily performed on a manually annotated dataset. A manually annotated dataset gives ground truth, however, manual labeling is an expensive and time-consuming task. In our proposed approach, we used active learning, a semi-supervised ML model that helps in classification when the amount of available labeled data is small. In this model, the classifier is trained with a small amount of training data (labeled data

TABLE II: Comparison of Models using Feature Selection

Papers	$R_S(u_i)$			h_Index		$C_S(u_i)$	$Twt_{cr}(u_i)$							URLs, List and Mentions			
	$F_{fol}(u_i)$	$F(u_i)$	$S(u_i)$	$R_{hindex}(u_i)$	$L_{hindex}(u_i)$		$R_R(u_i)$	$L_R(u_i)$	$H(u_i)$	$U(u_i)$	$O_R(u_i)$	$N_T(u_i)$	$M_R(u_i)$	$M(u_i)$	$U_R(u_i)$	$L(u_i)$
[47]	✓	✓				✓	✓					✓		✓	✓	
[31]	✓	✓				✓	✓	✓	✓	✓	✓		✓	✓	✓	
[41]	✓	✓				✓	✓	✓			✓	✓		✓	✓	
[54]	✓	✓				✓					✓		✓	✓	✓	
[55]	✓	✓				✓	✓		✓	✓		✓			✓	
[56]	✓	✓				✓	✓		✓	✓			✓	✓	✓	
Proposed	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

points). Then the points ambiguous to the classifier in the large pool of unlabeled data points are labeled and added to the training set [63]. This process is repeated until all the ambiguous instances are queried or the model performance does not improve above a certain threshold. Based on the proposed model, we train our classifier on a small human-annotated dataset which further classifies a large pool of unlabeled data points efficiently and accurately.

Our active learning process evolves through the following steps:

- **Data Gathering:** First, we gather the unlabeled data for 50,000 Twitter users. The unlabeled data is split into a seed – a small labeled dataset (manually labeled) and a large pool of unlabeled data. The seed is used to train the classifier just like a normal ML model. Using a dataset (seed) of 1,000 manually annotated data, we classify each political Twitter user as a trustworthy or untrustworthy user.
- **Selection of Unlabeled Instances:** A pool-based sampling with a batch size of 100 is used in which 100 ambiguous instances from the unlabeled dataset are labeled and added to a labeled dataset. Different sampling is employed to select the instances from the unlabeled dataset. For the new labeled dataset, the classifier is re-trained and then the next batch of ambiguous unlabeled instances for labeling are selected. The process is repeated until the model performance does not improve above a certain threshold.

In addition, we used the following two classifiers:

- **Random Forest Classifier (RFC):** RFC is an ensemble tree-based learning algorithm [65]. It aggregates the votes from different decision trees to decide the output class of the instance. RFC can run efficiently on large datasets, handle thousands of input variables, measure the relative importance of each feature, and produces a highly accurate classifier.
- **Support Vector Machine (SVM):** SVM produces high accuracy with less computation power and is widely used in classification tasks. To classify the instances, the SVM finds a hyperplane in N-dimensional space, where N represents the number of features [64]. The goal of SVM is to perform classification by finding the hyperplane separating the two classes more accurately (maximising the margin between two classes).

V. Experimental Results and Model Evaluation

Experimental Setup: To extract the features from Twitter and generate the dataset we used Python 3.5. The python script was executed locally on a machine with the following configuration: Intel Core i7, 2.80*8 GHZ, 32GB, Ubuntu 16.04 LTS 64 bit. For the training and evaluation of the machine learning models, we switched to Google Colab. We use the modAL framework [66], which is an active learning framework for python. It is a flexible, modular, and extensible framework built on top of Scikit learn. For the learner to query the instance labels, we use pool-based sampling and for the query strategy, we use different sampling techniques. For classification purposes, we use two RFC and SVC classifiers implemented using the scikit-learn library.

A. Dataset and Data Collection

To collect user features and tweets we used tweepy – Twitter's search API. Tweepy has certain limitations, as it only allows the collection of a certain number of parameters. Additionally, there is also a data rate limit that prevents the collection of information above a certain threshold. In our dataset, we chose to analyse the Twitter accounts of 50,000 politicians.

The main reason we decided to evaluate the profiles of politicians is their intrinsic potential to influence public opinion as their content originates and exists in a sphere of political life which is, unfortunately, often surrounded by controversial events and outcomes. When selecting these politicians, we only considered those with a *public profile* while users that seemed to be *inactive* (e.g. limited number of followers and activities) were omitted. Finally, for each user we extracted all the necessary features required by our model.

Using the extracted features and tweets, we calculated an influence score for each user. Furthermore, we generated a dataset consisting of 19 features including the influence score for 50,000 Twitter users. There are features such as the number of followers, likes, etc. which have no defined upper limits and may have outlying values. Hence, for these features, we used a percentile clip. We then normalised our features using min-max normalisation, with 0 being the smallest and 1 being the largest value.

B. Performance Measurements of Machine Learning and Neural Network Models

We garnered 50,000 unlabeled instances of Twitter users. The dataset was divided into three sets: training, testing, and unlabeled pool data. For the training and testing cohorts, we had 1,000 data points that were manually annotated. The rest of the data was unlabeled (49,000 instances). For the classification, we used the RFC and SVM classifiers (both classifiers are trained on the labeled dataset). Accuracy (%) is

used as the evaluation metric for model performance, which measures the percentage of correctly classified instances. To improve model accuracy, the active learner randomly selects ambiguous data points from the unlabeled data pool using three different sampling techniques and a person manually annotates the selected data. The annotated data is then added to the labeled dataset. This process is repeated 100 times for both the classifiers. The respective sampling techniques and accuracy obtained for both classifiers are discussed below.

Uncertainty Sampling: In uncertainty sampling, the instance in which there is the least confidence is most likely to be considered. In this type of sampling method, the most probable labels are considered and the rest are discarded. The results are shown in figure 2a for the RFC, which achieves an accuracy of 97.6% while the SVM obtained an accuracy of 96.8% (this is shown in figure 2b).

Margin Sampling: In margin sampling, instances with the smallest difference between the first and second most probable labels are considered. The accuracy for RFC and SVM using margin sampling is 97.6% and 96.2%, respectively. This can be seen in figure 3a for RFC and figure 3b.

Entropy Sampling: Lastly, the entropy sampling method offers the best results, obtaining an accuracy of 98.4% for RFC and 97% for SVM. An explanation for this improved performance can be attributed to the fact that entropy sampling utilises all possible label probabilities, unlike the other sampling methods. For RFC and SVM, this is also shown in figure 4a and 4b.

Open Science & Reproducible Research: As a way to support open science and reproducible research and give the opportunity to other researchers to use, test and hopefully extend/enhance our models, we plan to make both our datasets as well as the code of our models available through the Zenodo research artifacts portal. This will not violate Twitter's developer terms. However, in order to maintain our anonymity, we will make this available in a camera-ready version if the paper is accepted.

VI. Conclusion

Having identified the significant impact of fake news on our lives, this work focused on finding ways to identify this kind of information and notify users about the possibility that a specific post from a Twitter user may not be credible. To do so, we designed a model that analyses Twitter users and assigns each a calculated score based on their social profiles, tweet credibility, sentiment score, and h-index score (retweets and likes). Users with a higher score are not only considered more influential but their tweets are also considered to have greater credibility. To achieve our goal, we generated a dataset of 50,000 Twitter users (politicians) along with a set of 19 features for each user. Then, we classified each Twitter user as trustworthy or untrustworthy using RFC and SVM classifiers. Moreover, we employed the active learner approach to label ambiguous unlabelled data points. During the evaluation of our models, we conducted extensive experiments using three sampling methods which show the effectiveness of our approach. We believe this work is an important step towards re-establishing user trust in social networks and a stepping stone towards building new bonds of trust between users.

We see this work as an important step towards engendering user trust in social networks and we believe that it can constitute the underpinnings for establishing trust relationships between users.

References

[1] M. Hindman and V. Barash, Disinformation, and Influence Campaigns on Twitter, 2018.

[2] A. Java, X. Song, T. Finin, and B. Tseng, Why we twitter: understanding microblogging usage and communities, in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 2007, pp. 5665.

[3] L. A. Adamic and N. Glance, The political blogosphere and the 2004 US election: divided they blog, in Proceedings of the 3rd international workshop on Link discovery, 2005, pp. 3643.

[4] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Gonalves, A. Flammini, and F. M. Menczer, Detecting and tracking political abuse in social media, in Fifth international AAAI conference on weblogs and social media, 2011.

[5] Tassos Dimitriou and Antonis Michalas. Multi-Party Trust Computation in Decentralized Environments. Proceedings of the 5th IFIP International Conference on New Technologies, Mobility & Security (NTMS12), Istanbul, Turkey, 2012.

[6] Tassos Dimitriou and Antonis Michalas. Multi-Party Trust Computation in Decentralized Environments in the Presence of Malicious Adversaries. Ad Hoc Networks Journal, a special issue on Smart Solutions for Mobility Supported Distributed and Embedded Systems, Elsevier, 2014.

[7] A. H. Wang, Dont follow me: Spam detection in twitter, in 2010 international conference on security and cryptography (SECRYPT), 2010, pp. 110.

[8] N. Jain, P. Agarwal, and J. Pruthi, HashJacker-detection and analysis of hashtag hijacking on Twitter, Int. J. Comput. Appl., vol. 114, no. 19, 2015.

[9] C. Grier, K. Thomas, V. Paxson, and M. Zhang, @ spam: the underground on 140 characters or less, in Proceedings of the 17th ACM conference on Computer and communications security, 2010, pp. 2737.

[10] H. Allcott and M. Gentzkow, Social media and fake news in the 2016 election, J. Econ. Perspect., vol. 31, no. 2, pp. 211236, 2017.

[11] E. Metzgar and A. Maruggi, Social media and the 2008 US presidential election., J. New Commun. Res., vol. 4, no. 1, 2009.

[12] Z. Saaya and T. W. Hong, THE DEVELOPMENT OF TRUST MATRIX FOR RECOGNIZING RELIABLE CONTENT IN SOCIAL MEDIA, Int. J. Comput., vol. 18, no. 1, pp. 6066, 2019.

[13] X. Zhou, A. Jain, V. V Phoha, and R. Zafarani, Fake News Early Detection: A Theory-driven Model, arXiv Prepr. arXiv1904.11679, 2019.

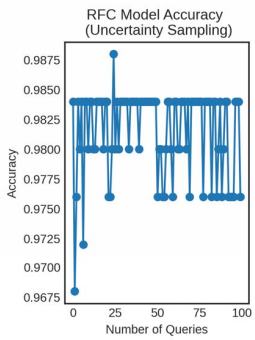
[14] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause, Fake news detection in social networks via crowd signals, in Companion Proceedings of the The Web Conference 2018, 2018, pp. 517524.

[15] A. Bovet and H. A. Makse, Influence of fake news in Twitter during the 2016 US presidential election, Nat. Commun., vol. 10, no. 1, p. 7, 2019.

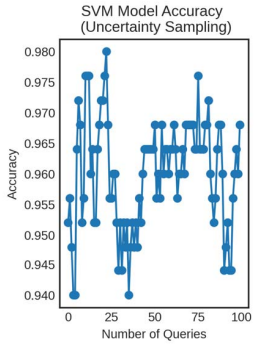
[16] M. Al-Qurishi, R. Aldrees, M. AlRubaian, M. Al-Rakhami, S. M. M. Rahman, and A. Alamri, A new model for classifying social media users according to their behaviors, in 2015 2nd World Symposium on Web Applications and Networking (WSWAN), 2015, pp. 15.

[17] K. R. Canini, B. Suh, and P. L. Pirolli, Finding credible information sources in social networks based on content and social structure, in 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 18.

[18] M. Gupta, P. Zhao, and J. Han, Evaluating event credibility on twitter, in Proceedings of the 2012 SIAM International Conference on Data Mining, 2012, pp. 153164.

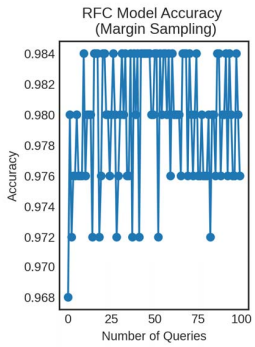


(a) RFC Model Accuracy Using Uncertainty Sampling

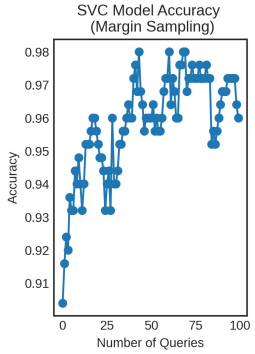


(b) SVM Model Accuracy Using Uncertainty Sampling

Fig. 2: Uncertainty Sampling

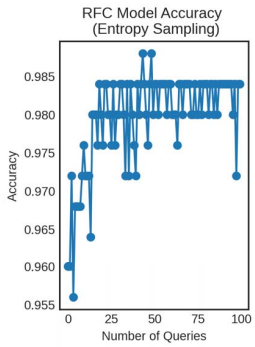


(a) RFC Model Accuracy Using Margin Sampling

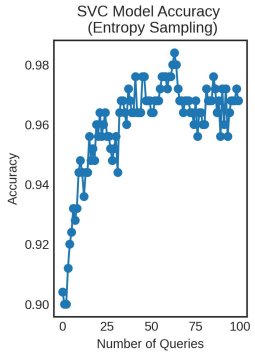


(b) SVM Model Accuracy Using Margin Sampling

Fig. 3: Margin Sampling



(a) RFC Model Accuracy Using Entropy Sampling



(b) SVM Model Accuracy Using Entropy Sampling

Fig. 4: Entropy Sampling

[19] F. Riquelme and P. Gonzlez-Cantergiani, Measuring user influence on Twitter: A survey, Inf. Process. & Manag., vol. 52, no. 5, pp. 949975, 2016.

[20] Y. Liu, C. Kliman-Silver, and A. Mislove, The tweets they are a-changin: Evolution of Twitter users and behavior, in Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[21] R. Tinati, L. Carr, W. Hall, and J. Bentwood, Identifying communicator roles in twitter, in Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 11611168.

[22] M.-F. Moens, J. Li, and T.-S. Chua, Mining user generated content. Chapman and Hall/CRC, 2014.

[23] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, Classifying latent user attributes in twitter, in Proceedings of the 2nd international workshop on Search and mining user-generated contents, 2010, pp. 3744.

[24] M. M. Uddin, M. Imran, and H. Sajjad, Understanding types of users on Twitter, arXiv Prepr. arXiv1406.1335, 2014.

[25] H. S. Al-Khalifa and R. M. Al-Eidan, An experimental system for measuring the credibility of news content in Twitter, Int. J. Web Inf. Syst., vol. 7, no. 2, pp. 130151, 2011.

[26] M. S. Granovetter, The strength of weak ties, in Social networks,

Elsevier, 1977, pp. 347367.
[27] I. Anger and C. Kittl, Measuring influence on Twitter, in Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, 2011, p. 31.
[28] H. S. Dutta, A. Chetan, B. Joshi, and T. Chakraborty, Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services, in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018, pp. 242249.
[29] M. Grandjean, A social network analysis of Twitter: Mapping the digital humanities community, Cogent Arts & Humanit., vol. 3, no. 1, p. 1171458, 2016.
[30] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 2236, 2017.
[31] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, Tweetcred: Real-time credibility assessment of content on twitter, in International Conference on Social Informatics, 2014, pp. 228243.
[32] M. Mendoza, B. Poblete, and C. Castillo, Twitter under crisis: Can we trust what we RT?, in Proceedings of the first workshop on social media analytics, 2010, pp. 7179.
[33] A. Gupta, H. Lamba, and P. Kumaraguru, 1.00 per rt bostonmarathon prayforboston: Analyzing fake content on twitter, in 2013 APWG eCrime researchers summit, 2013, pp. 112.
[34] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy, in Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 729736.
[35] C. Castillo, M. Mendoza, and B. Poblete, Information credibility on twitter, in Proceedings of the 20th international conference on World wide web, 2011, pp. 675684.
[36] A. Gupta and P. Kumaraguru, Twitter explodes with activity in mumbai blasts! a lifeline or an unmonitored daemon in the lurking?, 2012.
[37] A. Gupta and P. Kumaraguru, Credibility ranking of tweets during high impact events, in Proceedings of the 1st workshop on privacy and security in online social media, 2012, p. 2.
[38] O. Oh, M. Agrawal, and H. R. Rao, Information control and terrorism: Tracking the Mumbai terrorist attack through twitter, Inf. Syst. Front., vol. 13, no. 1, pp. 3343, 2011.
[39] K. Lee, J. Mahmud, J. Chen, M. Zhou, and J. Nichols, Who will retweet this?: Automatically identifying and engaging strangers on twitter to spread information, in Proceedings of the 19th international conference on Intelligent User Interfaces, 2014, pp. 247256.
[40] S. Lee and J. Kim, Warningbird: A near real-time detection system for suspicious urls in twitter stream, IEEE Trans. dependable Secur. Comput., vol. 10, no. 3, pp. 183195, 2013.
[41] Z. Gilani, R. Farahbakhsh, G. Tyson, L. Wang, and J. Crowcroft, An in-depth characterisation of Bots and Humans on Twitter, arXiv Prepr. arXiv1704.01508, 2017.
[42] Z. Gilani, E. Kochmar, and J. Crowcroft, Classification of twitter accounts into automated agents and human users, in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 2017, pp. 489496.
[43] Z. Gilani, R. Farahbakhsh, G. Tyson, L. Wang, and J. Crowcroft, Of bots and humans (on twitter), in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 2017, pp. 349354.
[44] S. Loria et al., Textblob: simplified text processing, Second. TextBlob Simpl. Text Process., 2014.
[45] G. Wolfsfeld, E. Segev, and T. Sheafer, Social media and the Arab Spring: Politics comes first, Int. J. Press., vol. 18, no. 2, pp. 115137, 2013.
[46] E. Morozov and M. Sen, Analysing the Twitter social graph: Whom can we trust?, MS thesis, Dept. Comput. Sci., Univ. Nice Sophia Antipolis, Nice, France, 2014.
[47] M. Alrubaian, M. Al-Qurishi, M. Al-Rakhami, M. M. Hassan, and A. Alamri, Reputation-based credibility analysis of Twitter social network users, Concurr. Comput. Pract. Exp., vol. 29, no. 7, p. e3873, 2017.
[48] A. L. Hughes and L. Palen, Twitter adoption and use in mass convergence and emergency events, Int. J. Emerg. Manag., vol. 6, no. 34, pp. 248260, 2009.
[49] X. Han, X. Gu, and S. Peng, Analysis of Tweet Forms effect on users engagement on Twitter, Cogent Bus. Manag., vol. 6, no. 1, pp. 115, Jan. 2019.
[50] B. Kang, J. ODonovan, and T. Hllerer, Modeling topic specific credibility on twitter, in Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, 2012, pp. 179188.
[51] J. ODonovan, B. Kang, G. Meyer, T. Hllerer, and S. Adalii, Credibility in context: An analysis of feature distributions in twitter, in 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, 2012, pp. 293301.
[52] D. Garcia, P. Mavrodiev, D. Casati, and F. Schweitzer, Understanding popularity, reputation, and social influence in the twitter society, Policy & Internet, vol. 9, no. 3, pp. 343364, 2017.
[53] G. M. Chen, Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others, Comput. Human Behav., vol. 27, no. 2, pp. 755762, 2011.
[54] A. A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, and C. Yang, Cats: Characterizing automation of twitter spammers, in 2013 Fifth International Conference on Communication Systems and Networks (COMSNETS), 2013, pp. 110.
[55] C. Yang, R. Harkreader, and G. Gu, Empirical evaluation and new design for fighting evolving twitter spammers, IEEE Trans. Inf. Forensics Secur., vol. 8, no. 8, pp. 12801293, 2013.
[56] M. Fazil and M. Abulaish, A hybrid approach for detecting automated spammers in twitter, IEEE Trans. Inf. Forensics Secur., vol. 13, no. 11, pp. 27072719, 2018.
[57] C. G. McCoy, M. L. Nelson, and M. C. Weigle, University Twitter engagement: using Twitter followers to rank universities, arXiv Prepr. arXiv1708.05790, 2017.
[58] A. Leavitt, E. Burchard, D. Fisher, and S. Gilbert, The influentials: New approaches for analyzing influence on twitter, Web Ecol. Proj., vol. 4, no. 2, pp. 118, 2009.
[59] A. Preussler and M. Kerres, Managing reputation by generating followers on Twitter, Medien. Explor. Vis. und kollaborativer Wissensrume, pp. 129143, 2010.
[60] M. Kerres and A. Preussler, Managing reputation by generating followers on Twitter. Medien-Wissen-Bildung, 2010.
[61] J. Clement, Ed., Number of social network users worldwide from 2010 to 2021 (in billions). 2019.
[62] A Beginners Guide to Active Learning - DataCamp. [Online]. Available: https://www.datacamp.com/community/tutorials/active-learning. [Accessed: 28-Jul-2020].
[63] B. Settles, Computer Sciences Department Active Learning Literature Survey, University of Wisconsin-Madison Department of Computer Sciences, 2009.
[64] W. S. Noble, What is a support vector machine?, Nature Biotechnology, vol. 24, no. 12. Nature Publishing Group, pp. 15651567, Dec-2006.
[65] G. Biau and G. B. Fr, Analysis of a Random Forests Model, 2012.
[66] T. Danka and P. Horvath, modAL: A modular active learning framework for Python, May 2018.
[67] D. K. Simon Tong, Support vector machine active learning with applications to text classification, J. Mach. Learn. Res., vol. 1, 2000.